ABSTRACT OF THE DISCLOSURE

Methods and apparatus are described for intelligently assigning a portion of a cluster's traffic (e.g., buckets) to a cache system to minimize overloading of such cache system. In general terms, when a new cache system enters a cache cluster and/or starts up, the new cache system's full bucket allocation is not immediately assigned to the new cache system. Instead, only a portion of the full bucket allocation is initially assigned to the new cache system. In one embodiment, the new cache system's bucket assignment is gradually increased until the cache system is handling it's full bucket allocation or it becomes overloaded. The cache system's load is also checked periodically to determine whether it has become overloaded. When the cache system becomes overloaded, buckets are immediately shed from the cache system. In sum, the new cache system's load is adjusted until it is handling an optimum number of buckets.